

美图大数据开放平台

陈满意 技术经理 2017.11.04



> 美图数据统计发展

美图大数据开放平台

> 未来规划

美图公司与大数据统计

10亿级用户

千万级日活 亿级月活

日保存图片2亿 日播放数亿级 日消息百亿级

PB级总数据 TB级日增量



美图数据统计发展历程

从无到有

• 简单实现、快速上线





业务与数据的爆发

业务爆发

数据量爆发

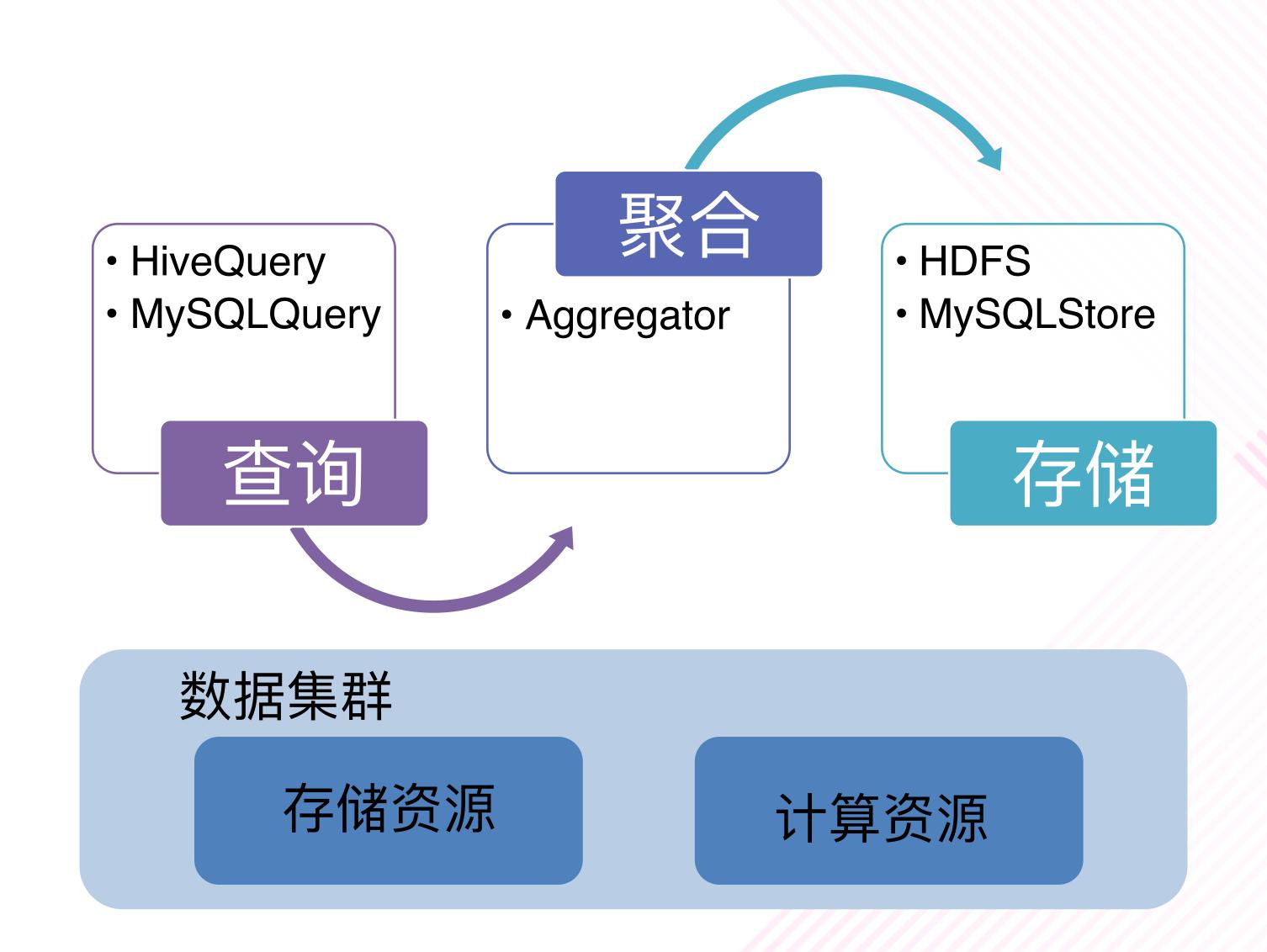
需求量膨胀

存储容量

计算资源

重复造轮子

进化中的统计





还有痛点

求量持续激增

计业务多样性

护成本高

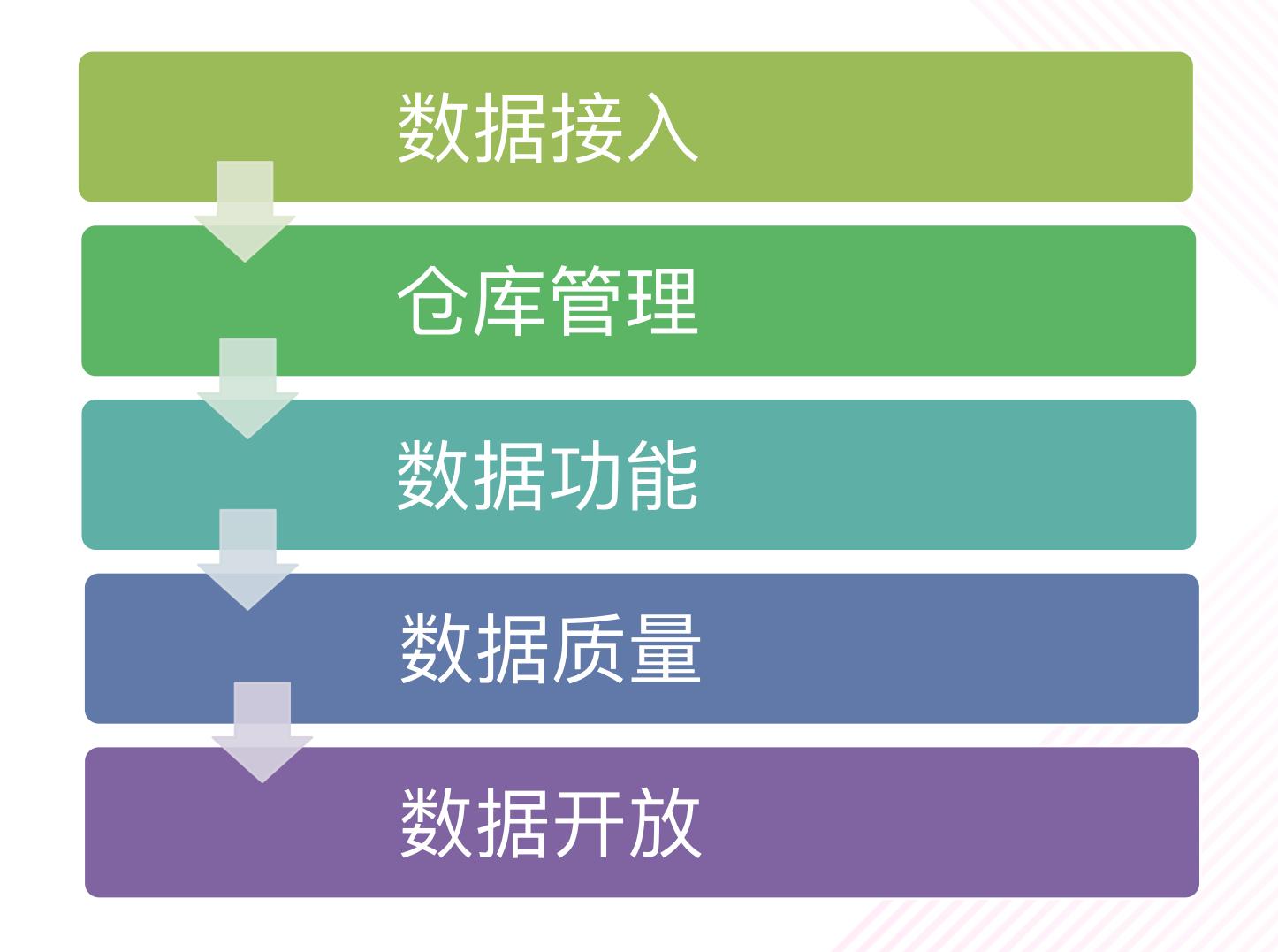


大数据开放平台



数据开放平台

• 美图大数据体系中提供数据服务的开放平台



数

据

开

放

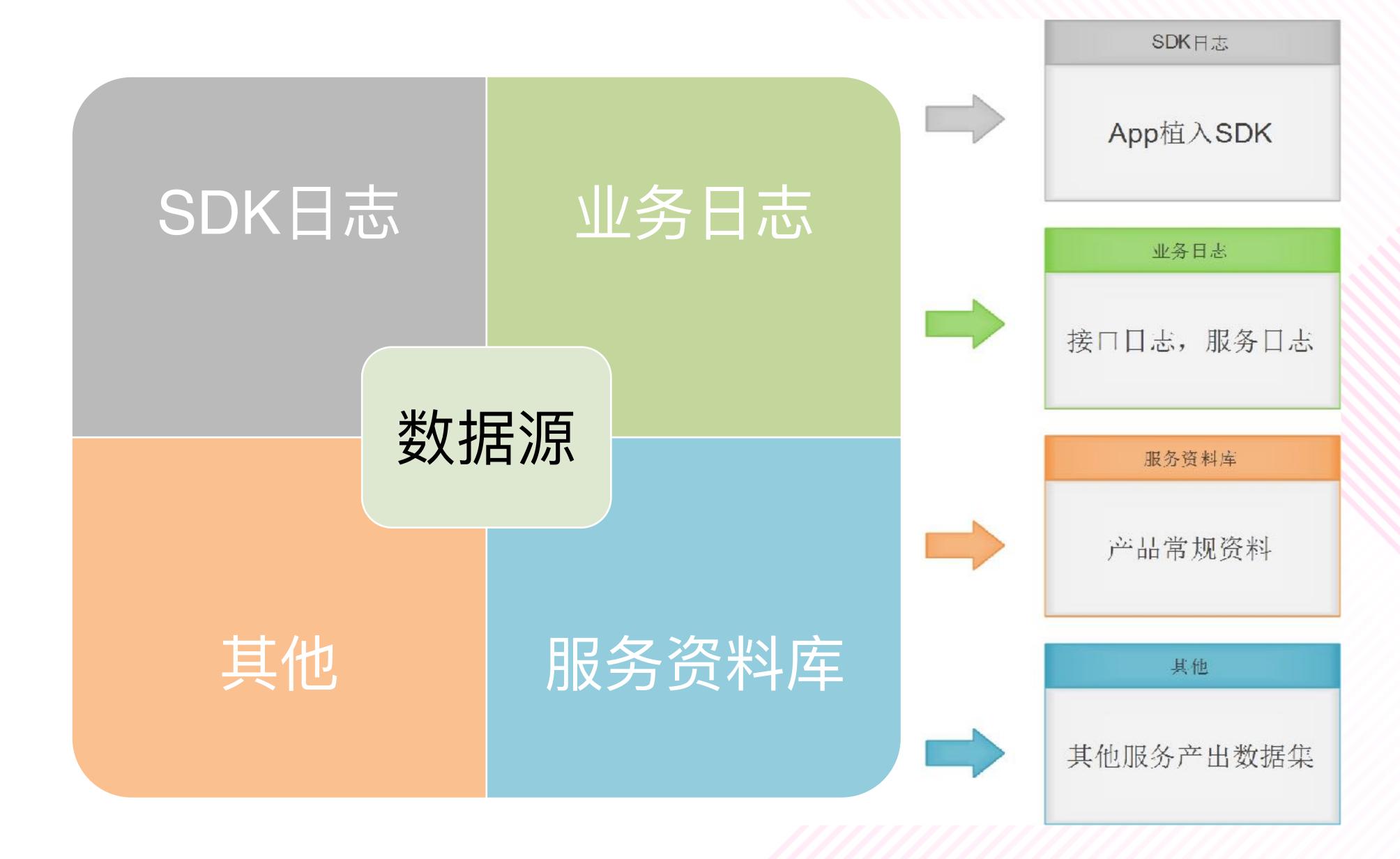
平台架构

Job 仓库建设 数据操作 同步分发 平台监控 元数据管理 数据质量 Manager 简洁方式 操作 原生操作 再封装 计数 去重 TopN 插件 Sqoop Hive **HDFS** Mongo Mysql File 分布式调度器

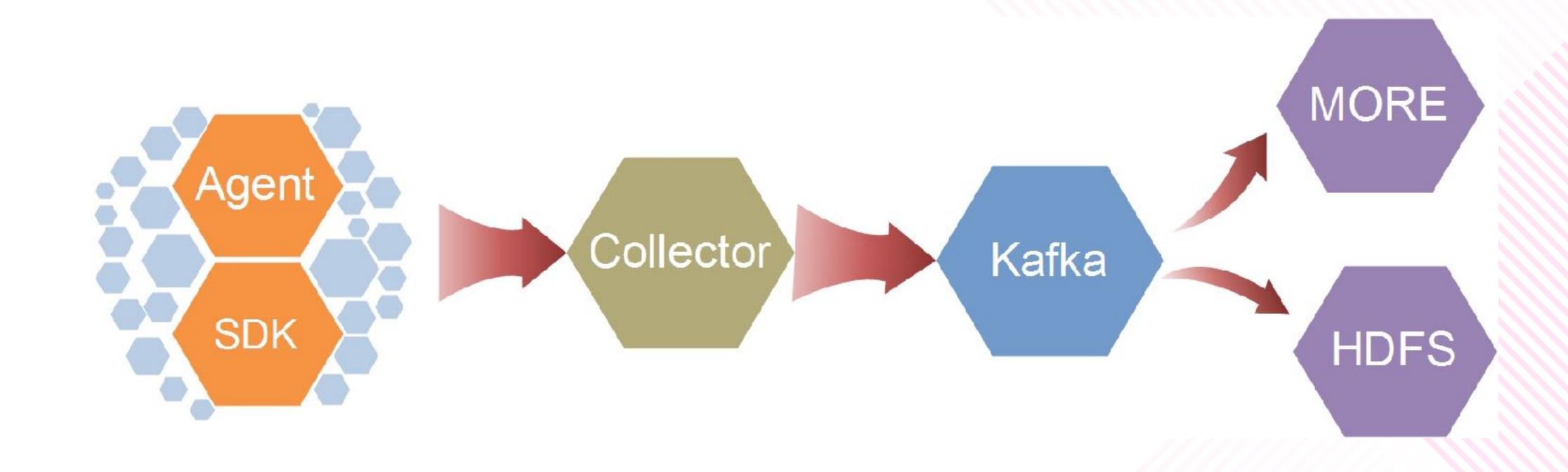
数据接入

数据接入-来源

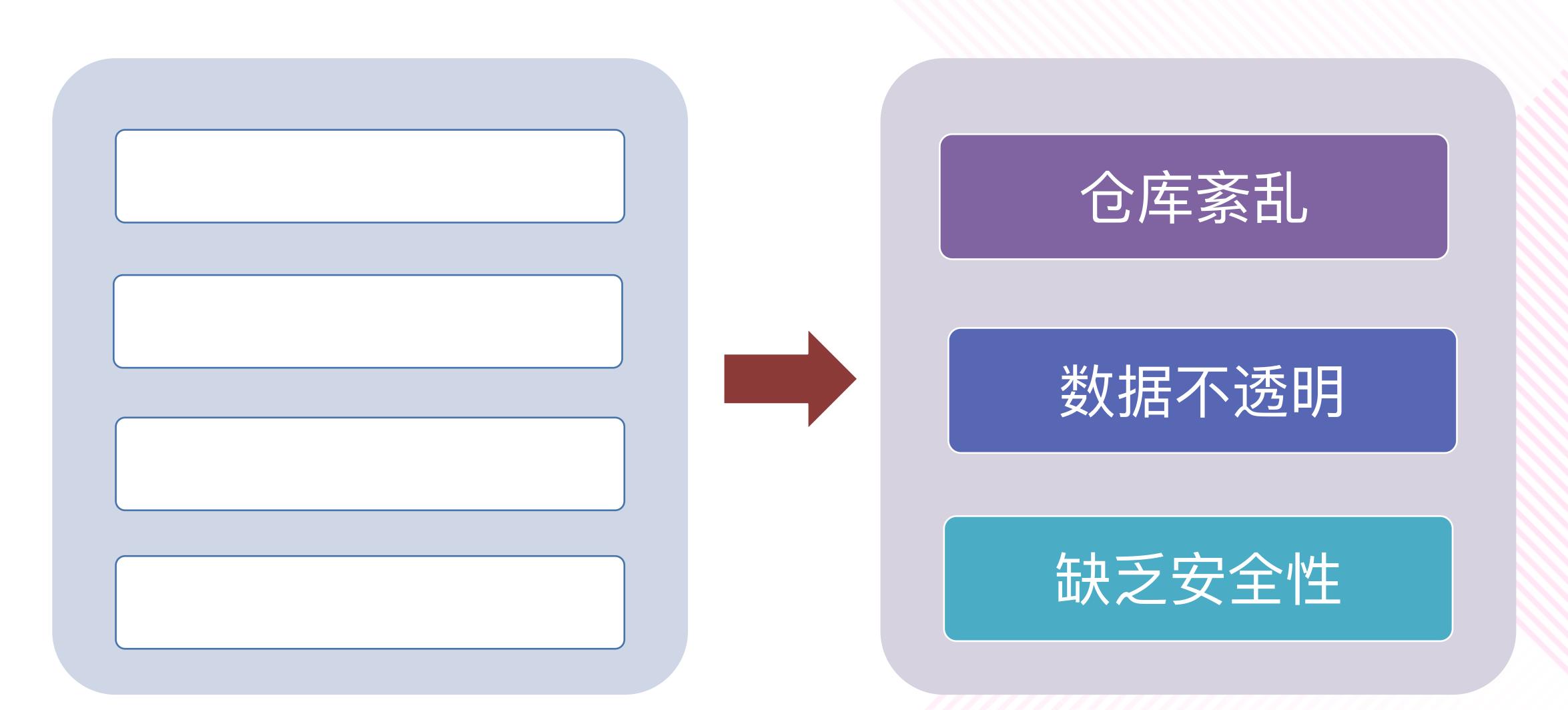




数据接入一流程

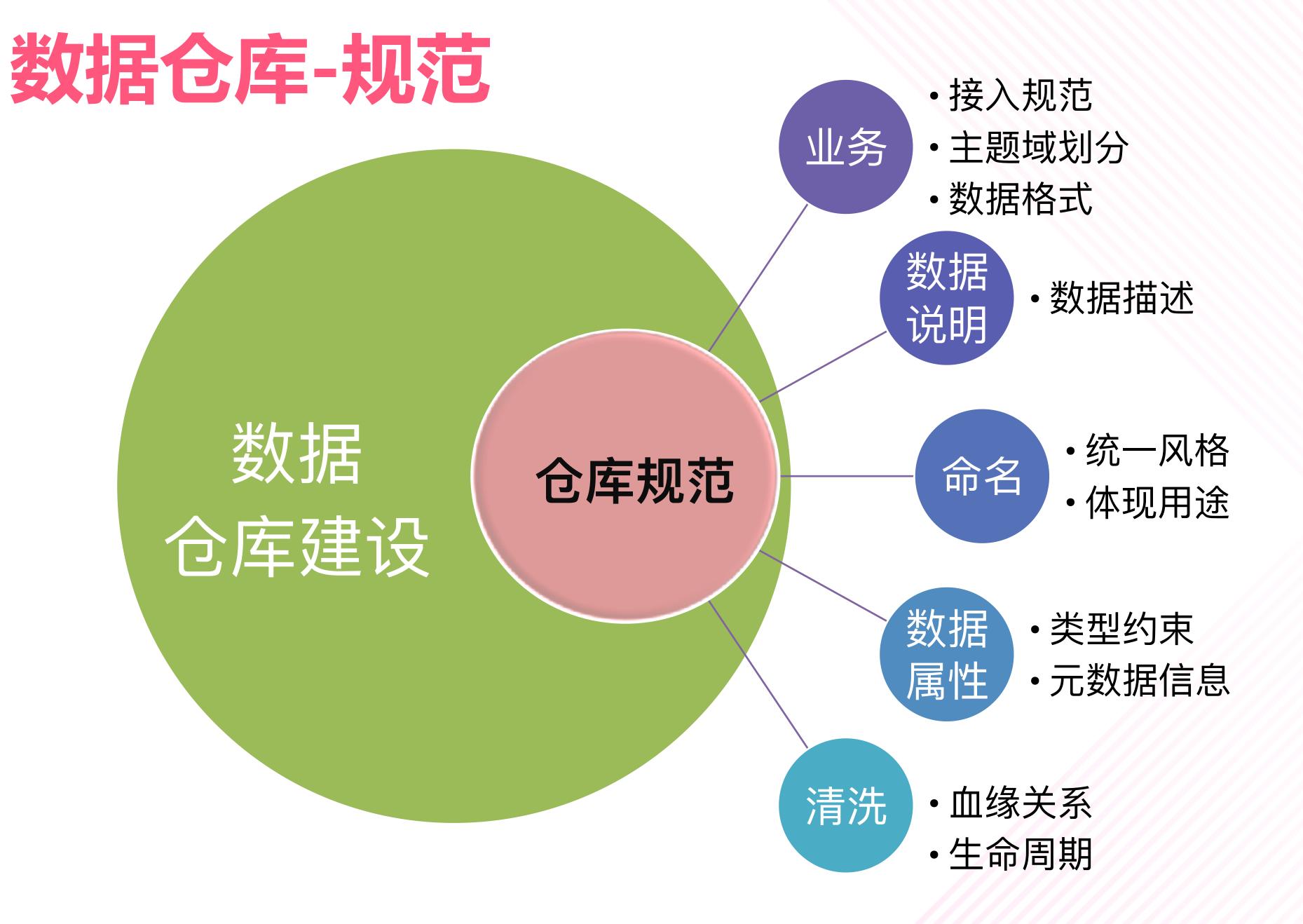


失控的HDFS与Hive



数据仓库-建设





数据查询

数据统计

- •丰富的统计插件
- 快速自定义插件
- 支持UDF

业务自治

- •非研发人员快速接入
- •可视化的任务配置
- ·计数\去重\TopN等
- •自动化生成接口

数据分析

- •自助查询
- 语法校验



数据查询-不受控的语法

级错误语法

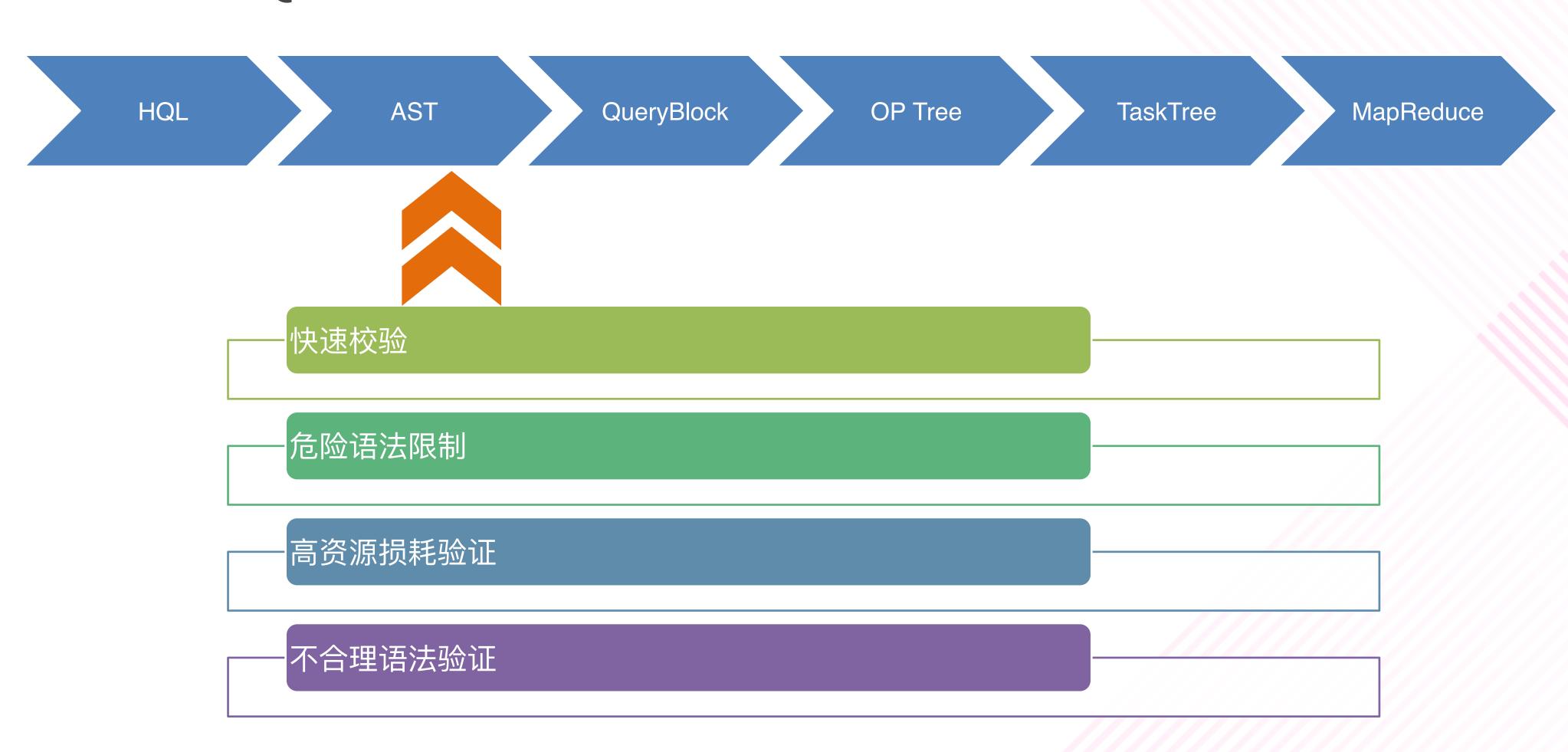
险操作语法

资源损耗的语法

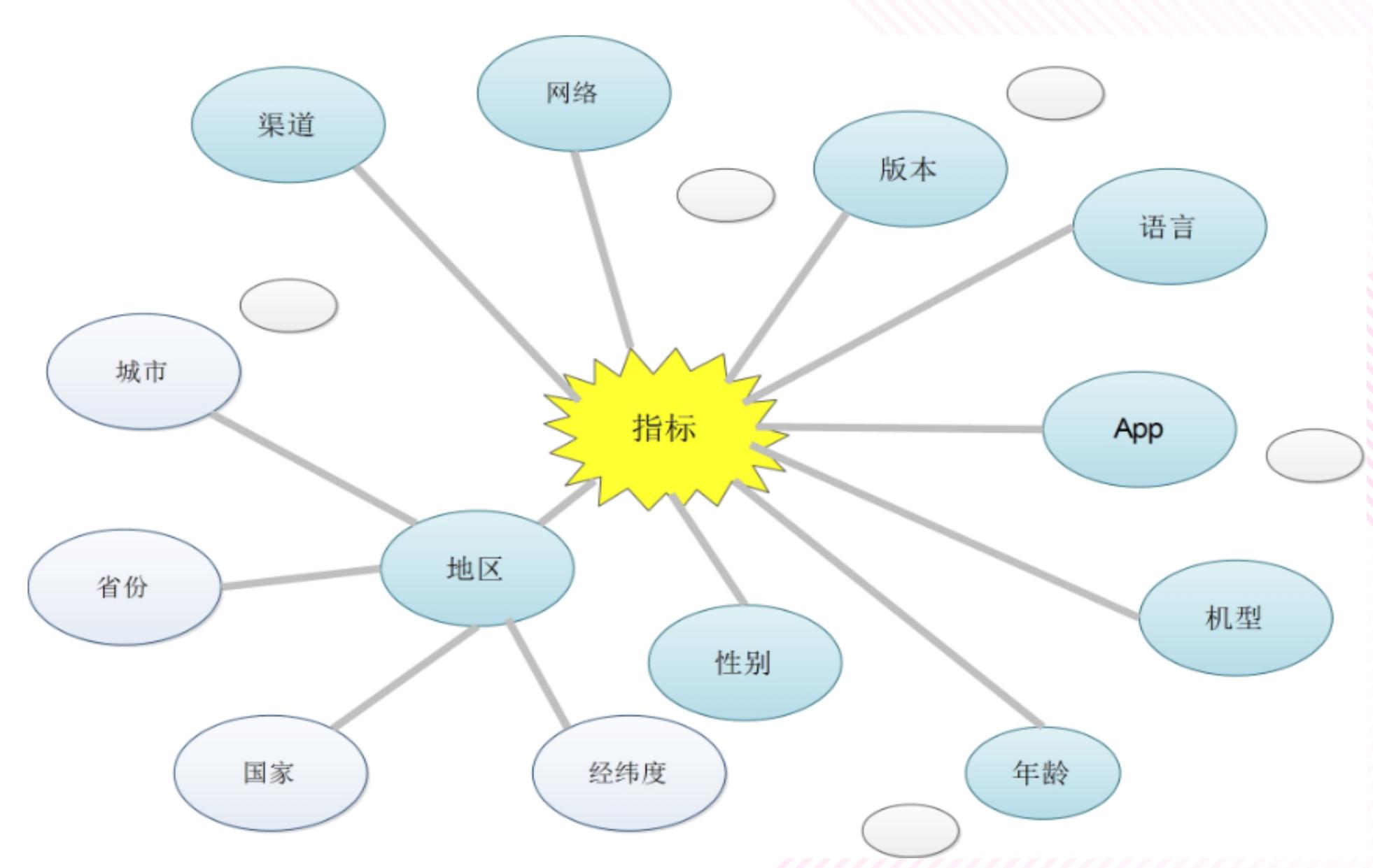


数据查询-语法验证

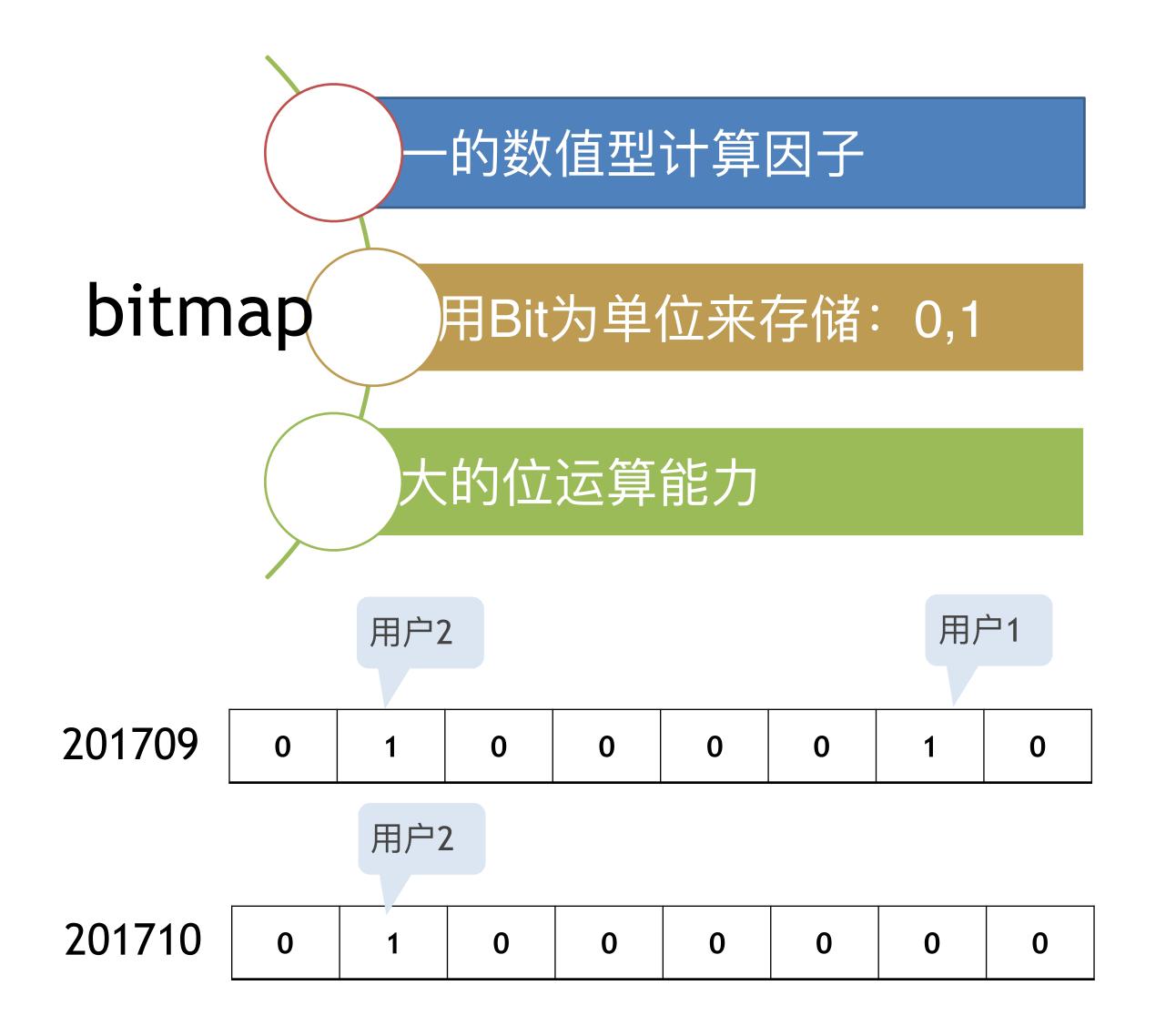
·基于HQL的语法验证

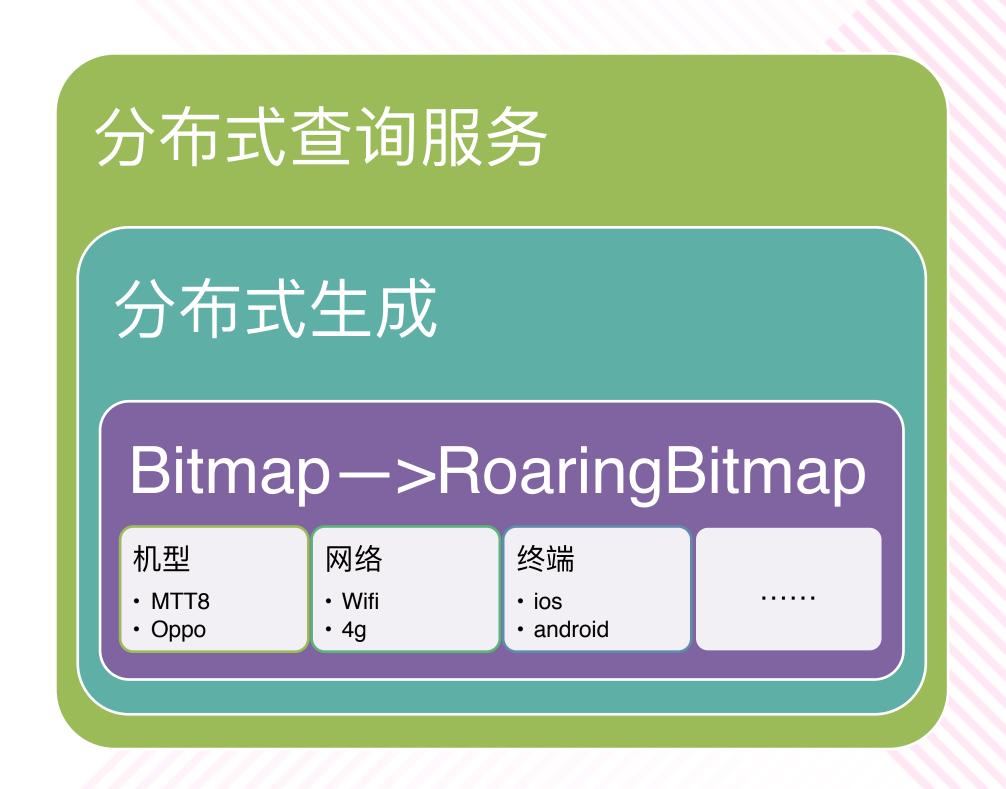


数据查询-多维度查询

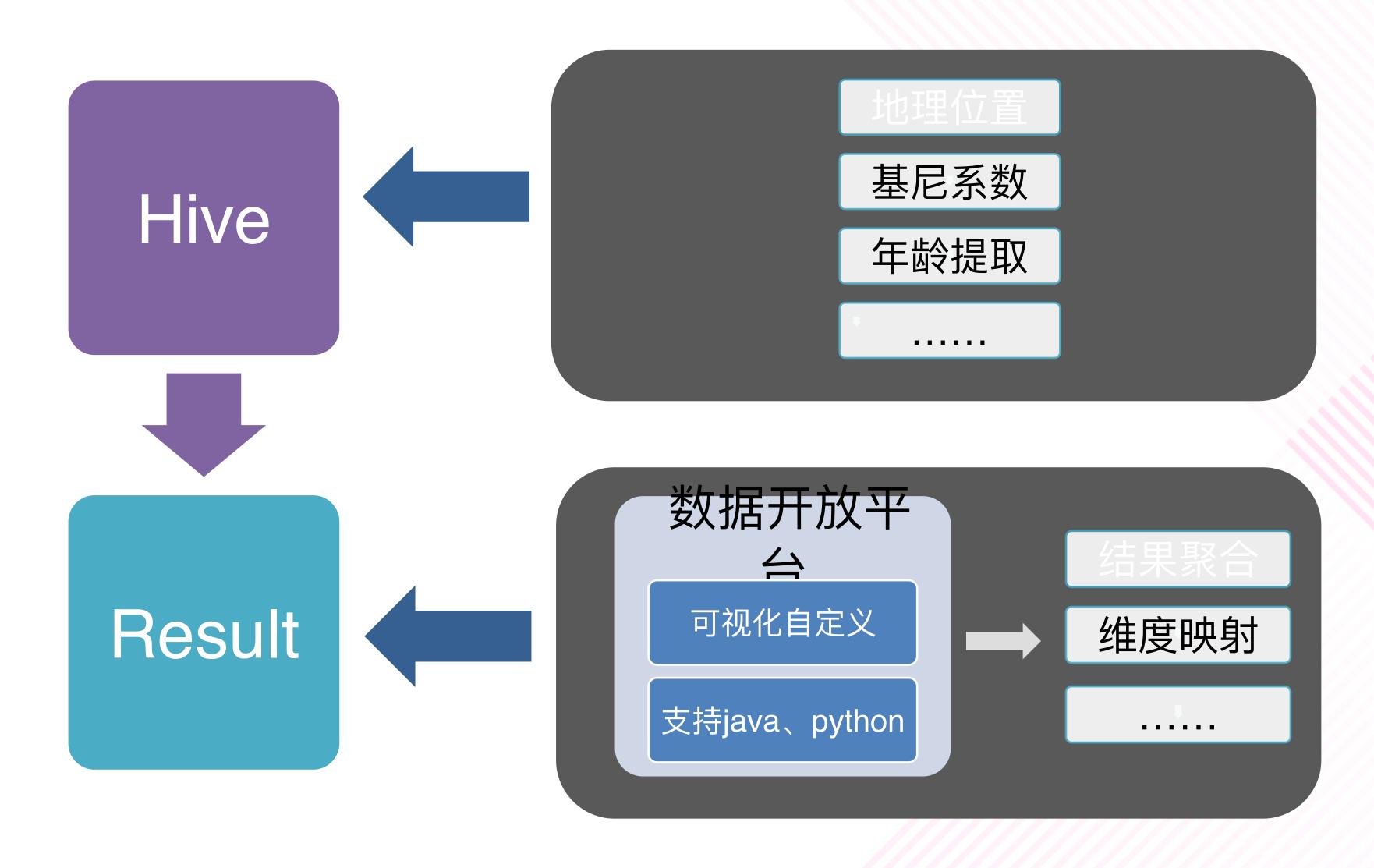


数据查询-多维度查询





自定义UDF



统计数据存储方案

mongodb

模式自由

•简化任务创建

快速

•近期指标为热数据

可靠性

•副本集

可扩展

•分片机制

作业调度器

• 作业是运行在分布式作业调度器上

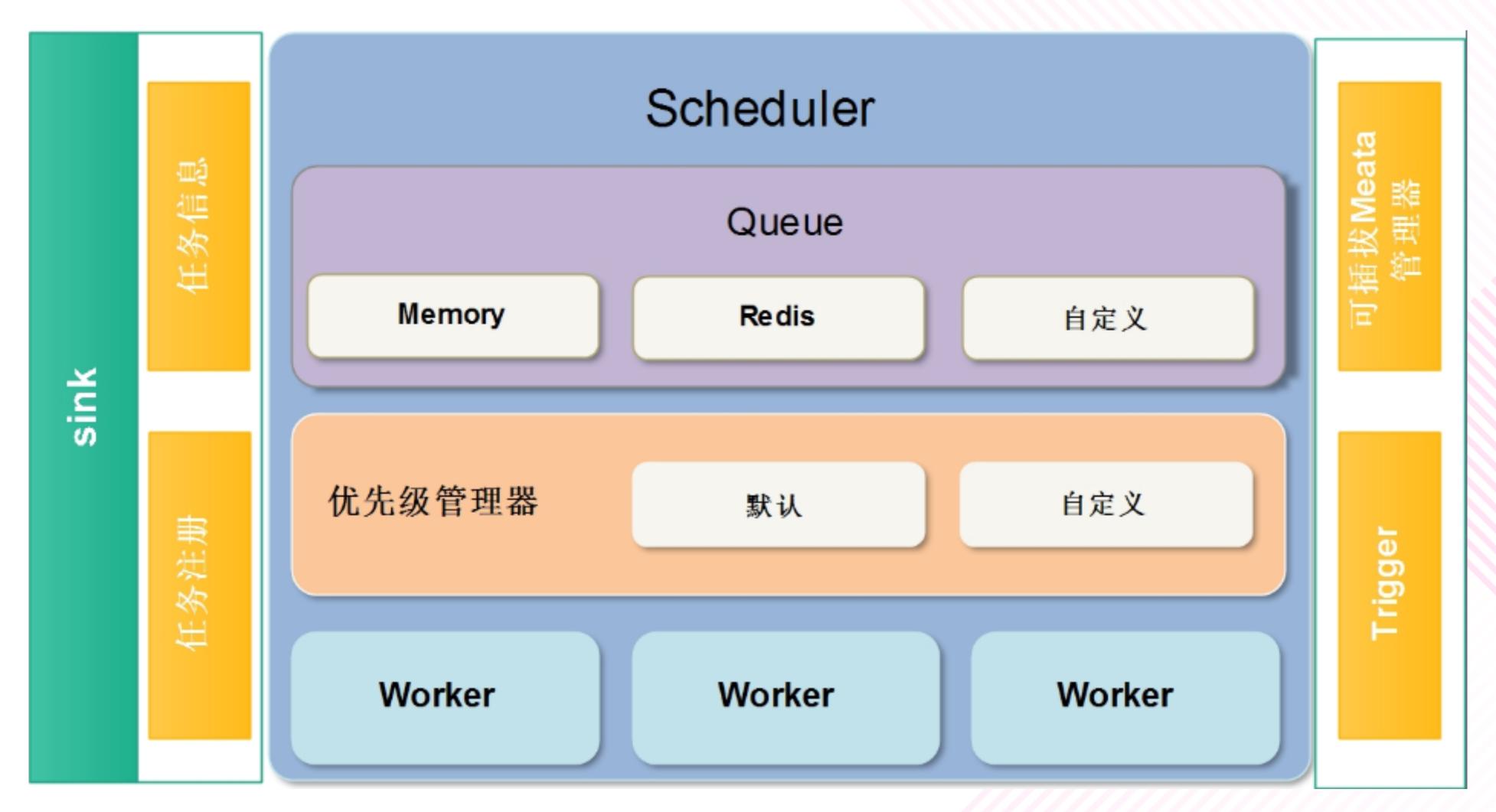
持即时/定时的作业调度

持工作流模式

持作业优先级等参数设置

持调度器横向扩展

作业调度器



数据质量

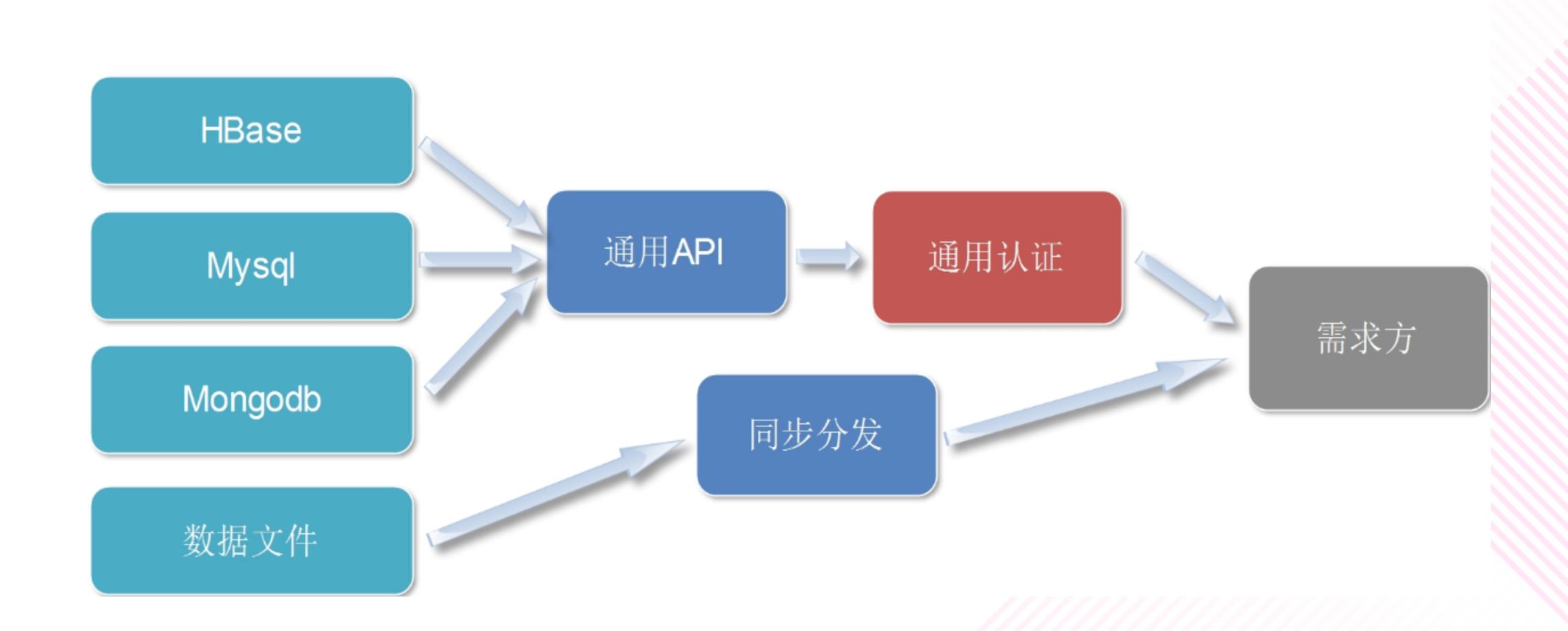


·统一的唯一设备ID

•数据的拓展服务

•数据监控服务

数据开放



美图大数据开放平台



